Structure of rat procathepsin B: model for inhibition of cysteine protease activity by the proregion

Miroslaw Cygler[1,2,3]*, J Sivaraman[1,2], Pawel Grochulski[1,3,4], René Coulombe[1,2], Andrew C Storer[1,2] and John S Mort[2,5]

Background: Cysteine proteases of the papain superfamily are synthesized as inactive precursors with a 60–110 residue N-terminal prosegment. The propeptides are potent inhibitors of their parent proteases. Although the proregion binding mode has been elucidated for all other protease classes, that of the cysteine proteases remained elusive.

Results: We report the three-dimensional structure of rat procathepsin B, determined at 2.8 Å resolution. The 62-residue proregion does not form a globular structure on its own, but folds along the surface of mature cathepsin B. The N-terminal part of the proregion packs against a surface loop, with Trp24p (p indicating the proregion) playing a pivotal role in these interactions. Inhibition occurs by blocking access to the active site: part of the proregion enters the substrate-binding cleft in a similar manner to a natural substrate, but in a reverse orientation.

Conclusions: The structure of procathepsin B provides the first insight into the mode of interaction between a mature cysteine protease from the papain superfamily and its prosegment. Maturation results in only one loop of cathepsin B changing conformation significantly, replacing contacts lost by removal of the prosegment. Contrary to many other proproteases, no rearrangement of the N terminus occurs following activation. Binding of the prosegment involves interactions with regions of the enzyme remote from the substrate-binding cleft and suggests a novel strategy for inhibitor design. The region of the prosegment where the activating cleavage occurs makes little contact with the enzyme, leading to speculation on the activation mechanism.

Addresses: [1]Biotechnology Research Institute, National Research Council of Canada, 6100 Royalmount Avenue, Montreal, Quebec H4P 2R2, Canada; [2]Protein Engineering Network of Centers of Excellence, Canada; [3]Montreal Joint Centre for Structural Biology, Montreal, Canada; [4]Institute of Physics, Technical University of Lódź, 93-005, Poland and [5] Shriners Hospital for Crippled Children and Department of Surgery, McGill University, Montreal, Quebec H3G 1A6, Canada.

*Corresponding author.

Key words: cathepsin B, cysteine protease, procathepsin B, proenzyme, three-dimensional structure

Received: **8 Jan 1996**
Revisions requested: **23 Jan 1996**
Revisions received: **5 Feb 1996**
Accepted: **6 Feb 1996**

Structure 15 April 1996, **4**:405–416

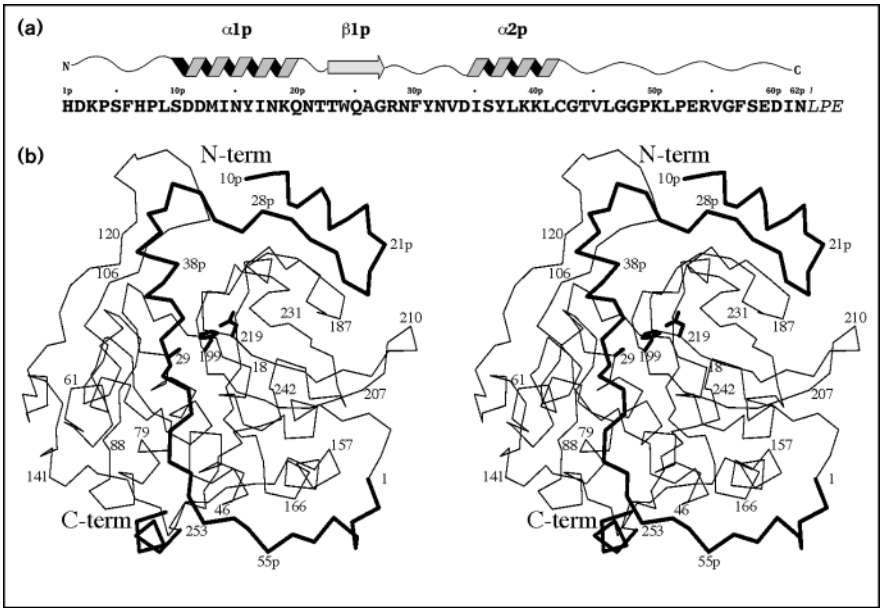
Introduction

Cysteine proteases represent a major component of the lysosomal proteolytic repertoire and play an important role in intracellular protein degradation [1]. They have also been implicated in proteolysis occurring within the endosomal system, in particular during antigen processing [2] and are thought to play an extracellular role in several physiological and pathological conditions, notably bone resorption [3], tumor invasion and metastasis [4] and cartilage degradation in arthritis [5]. The role of cysteine proteases in pathological conditions makes them important targets for inhibitor development. Although highly potent inhibitors are known, their major drawback is a lack of sufficient specificity to allow the targeting of a particular cysteine protease.

As is the case with essentially all proteases [6], the lysosomal cysteine proteases are produced as inactive precursors due to the presence of an additional N-terminal segment that has been shown to act as a potent, pH-dependent inhibitor ([7,8]; E Carmona, *et al.*, & R Ménard, unpublished data). The propeptide has also been shown to be essential for the correct folding of some cysteine proteases [9] and for their stabilization upon exposure to neutral or alkaline pH environments [10]. A similar chaperone role of the propeptide in the folding of the enzyme has been observed for serine proteases [11]. In the case of cysteine proteases, cleavage and dissociation of the propeptide with concomitant activation of the protease occurs as the result of autoprocessing under acidic conditions [12–14]. *In vivo* this takes place as the newly synthesized proenzyme is routed from the Golgi apparatus to the lysosome. It is not yet clear if the processing *in vivo* is an intra- or intermolecular event although processing *in vitro* can be achieved by intermolecular events [14,15].

Cathepsin B (catB) is one of the most studied of the lysosomal cysteine proteases. Unlike most other enzymes of this family it exhibits both endopeptidase and exopeptidase activities. In addition to cleavage within the peptide substrate, catB can also remove dipeptide units from the C terminus (peptidyldipeptidase activity [16]). It has been shown recently that the exopeptidase activity is dependent on the presence of a specialized structural element, the occluding loop, which accepts the negative charge of the P_2' carboxylate (JSM, in preparation).

Figure 1



(a) The amino-acid sequence of prosegment of rat procathepsin B, residues 1p–62p, represented in one-letter code, followed by the beginning of the mature sequence (italics). Secondary structural elements are marked above the sequence. **(b)** Stereoview of the Cα trace of procathepsin B. The prosegment and C-terminal extension are shown in thicker lines. Catalytic residues Ser29 (Cys), His199 and Asn219 are shown in full. (Figure was prepared with the program MOLSCRIPT [47].)

CatB belongs to the papain superfamily and shows high homology to cathepsins L, S and O, papain and actinidin, among others [17]. This superfamily encompasses a large number of cysteine proteases from sources as diverse as bacteria, plants and mammals. Recently it has become clear that these enzymes can be divided into two subfamilies based, in part, on the length of their proregions [18]. Thus, while the majority of cysteine proteases have a prosegment on the order of 100 residues, catB and its homologs have a much shorter, ~60-residue prosegment. It has been shown for catB that the prosegment is a powerful inhibitor of the enzyme's proteolytic activity and that this inhibition is highly specific towards catB, with a much lower activity against other related cysteine proteases [7].

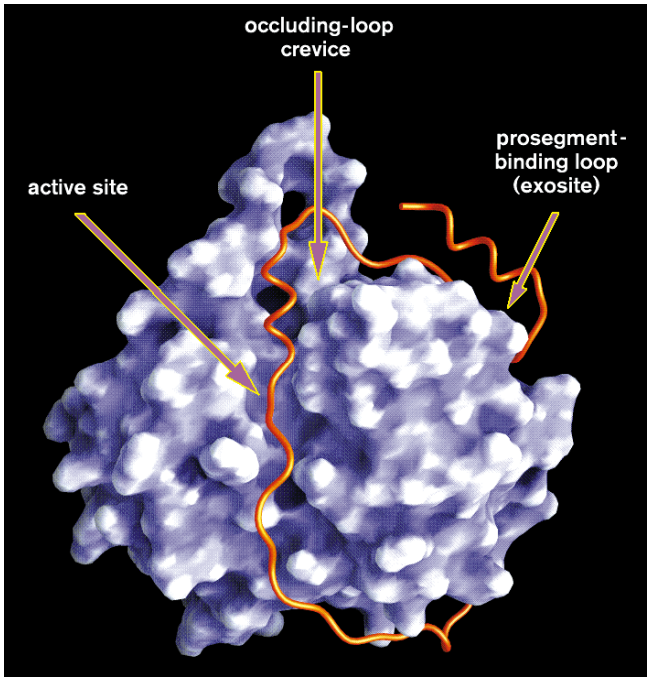
Crystallographic studies have demonstrated the structural basis for the inactivation of serine [19,20], aspartic [21] and metalloproteases [22,23] by their prosegments and have underlined the diverse means by which the proenzymes achieve the same goal. For cysteine proteases in general, and for cysteine proteases from the papain superfamily in particular, such structural data are currently lacking. The length of the prosegment of catB, 62 residues, indicates that it could form a globular structure that folds independently of the rest of the enzyme. However, solution studies by NMR spectroscopy of synthetic peptides corresponding to the entire propeptide, as well as various shorter fragments of the proregion, have shown that these peptides possess very little ordered structure, but that some α-helical segments form upon binding to the mature cathepsin B (F Ni, personal communication).

The mature enzyme, catB, is structurally very well characterized [24]. The structures of this enzyme of human and rat origin complexed with various inhibitors have been determined [25,26] and the substrate-binding subsites have been mapped. Here we present the three-dimensional structure of recombinant, inactive mutant of rat procathepsin B at 2.8 Å resolution. The Cys29→Ser mutation was introduced to reduce the problems of autoprocessing and this mutant was crystallized. We describe the conformation of the proregion, delineate its interaction with the rest of the molecule and identify the potential determinants of its specificity.

Results and discussion

Throughout the text the residue number for the prosegment is followed by a lower case letter p and the following residues are numbered as in mature enzyme, for example the last residue of the prosegment is Asn62p and the catB sequence starts with Leu1. We will use the term proregion or prosegment interchangeably and will refer to the mature part of the proenzyme simply as catB. The sequence of the prosegment is presented in Figure 1a and a stereodiagram of procathepsin B is shown in Figure 1b. The view of the molecular surface of catB with the Cα trace of the prosegment is shown in Figure 2. The part of the proenzyme corresponding to catB, residues 1–254, is well defined in the electron-density map. Its fold is the same as that of the mature rat enzyme [26]. The structural differences are relatively small, with the exception of one region, the so called occluding loop (see below). The protein consists of two domains, one of them is composed mostly of β sheet and the other contains several α helices and a small β-sheet

Figure 2



Molecular surface of the mature portion of procathepsin B with the prosegment shown as a worm representation. The view is toward the active site. (The figure was prepared using GRASP [48].)

structure [27]. The active-site triad, Cys29–His199–Asn219, and the substrate-binding site are located at the interface between the two domains. The procathepsin B gene codes for a protein with six additional residues at the C terminus compared with catB. These residues are removed during the maturation process, together with the proregion. Their role, if any, in the normal functioning of procathepsin B is unknown. In the crystal structure they are well defined and form a somewhat distorted two-turn α helix extending along the direction of a preceding β strand.

The proregion

The first nine N-terminal residues of the proregion are not well defined in the electron-density maps in either of the two molecules in the asymmetric unit, although there is some scattered density in this region. Non-crystallographic symmetry averaging did not improve the density for these residues, indicating that the N terminus is disordered and mobile in the crystal. The remaining part of the proregion has no globular structure on its own (Fig. 3a), but folds over catB. In this respect, the prosegment is similar to the 44-residue-long prosegment of pepsinogen, which also has very little secondary structure and extends along the surface of the protein [21,28].

The N terminus of the proregion folds into a hairpin structure formed by a three-turn α helix (α1p; residues 11p–20p), a tight bend (21p–23p) and a strand of extended polypeptide chain (β1p; 24p–28p) running antiparallel to the helix (Fig. 3a). The small hydrophobic core of this supersecondary structural element is formed by two isoleucines, Ile14p and Ile17p, along one side of the helix, and Trp24p and Ala26p from the β1p strand (Fig. 3b). The aliphatic part of the Arg28p side chain also contributes to this hydrophobic core and packs against Ile14p. The electron density of this arginine is weak, especially at the tip, indicating a partial disorder with a possibility of a weak salt bridge with Asp11p. This hairpin binds to catB on the opposite side of the enzyme to the location of the N terminus of the mature enzyme (Fig. 1b).

The hairpin is connected, through a loop, to a short two-turn α helix (α2p; 35p–41p), which forms an anchor, attaching to the S′ side of the active site, that helps fasten the polypeptide chain to the substrate- binding cleft. The prosegment dips into the cleft and re-emerges on the S side. The propeptide is accommodated in the binding-site cleft very much like the complexes between enzymes and peptidyl inhibitors studied so far, which are thought to represent the substrate-binding mode. But, importantly the propeptide follows the cleft in a reverse direction to that of natural substrates (see below).

From the substrate-binding site the prosegment follows along the surface of the protein towards Leu1, the N-terminal residue of catB (Fig. 1b). The electron density corresponding to the last part of the proregion is less well defined than in other regions and some side chains have not been located in the map, indicating an increased mobility of this region, resulting from a limited number of interactions with catB.

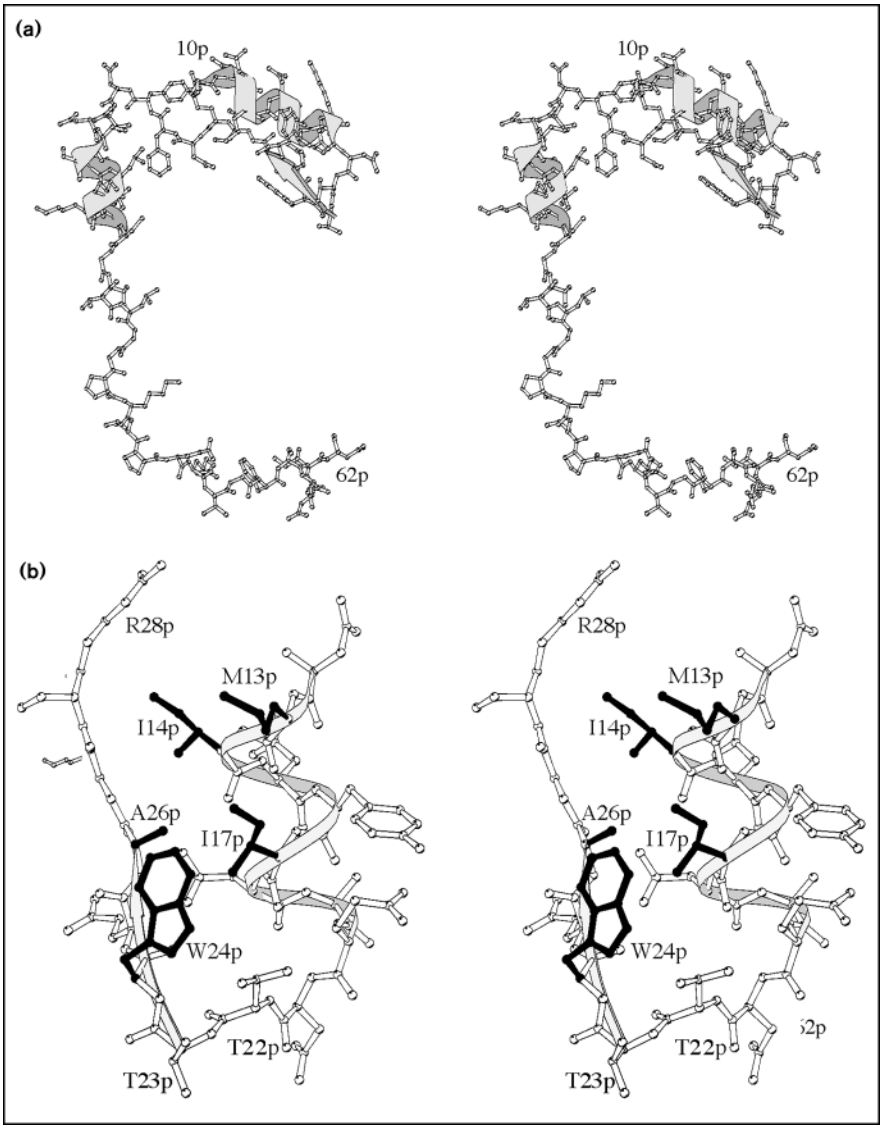
Proregion–catB interactions

With the exception of a disordered N terminus, almost the entire length of the proregion interacts with catB. Nevertheless, the majority of close contacts with catB are provided by residues 13p–47p. CatB residues in contact with the proregion are located within three major areas: a large 176–194 surface Ω-loop (prosegment-binding loop or PBL), a crevice between the occluding loop (108–122) and one side of PBL, and the substrate-binding cleft. The association of the proregion with catB involves a large contact area: 1950 Å^2 of the proregion surface and 2090 Å^2 of the catB surface is buried in the complex.

Prosegment-binding loop

The main interactions of the prosegment occur with residues from the PBL. The center of this loop, occupied by two aromatic side chains, Tyr183 and Tyr188, forms a depression on the surface of catB. These aromatic residues are part of a larger aromatic cluster that contains, in addition, Phe174, Phe180, Trp221, Trp225 and Phe231, forming a typical herring-bone-like pattern. The

Figure 3

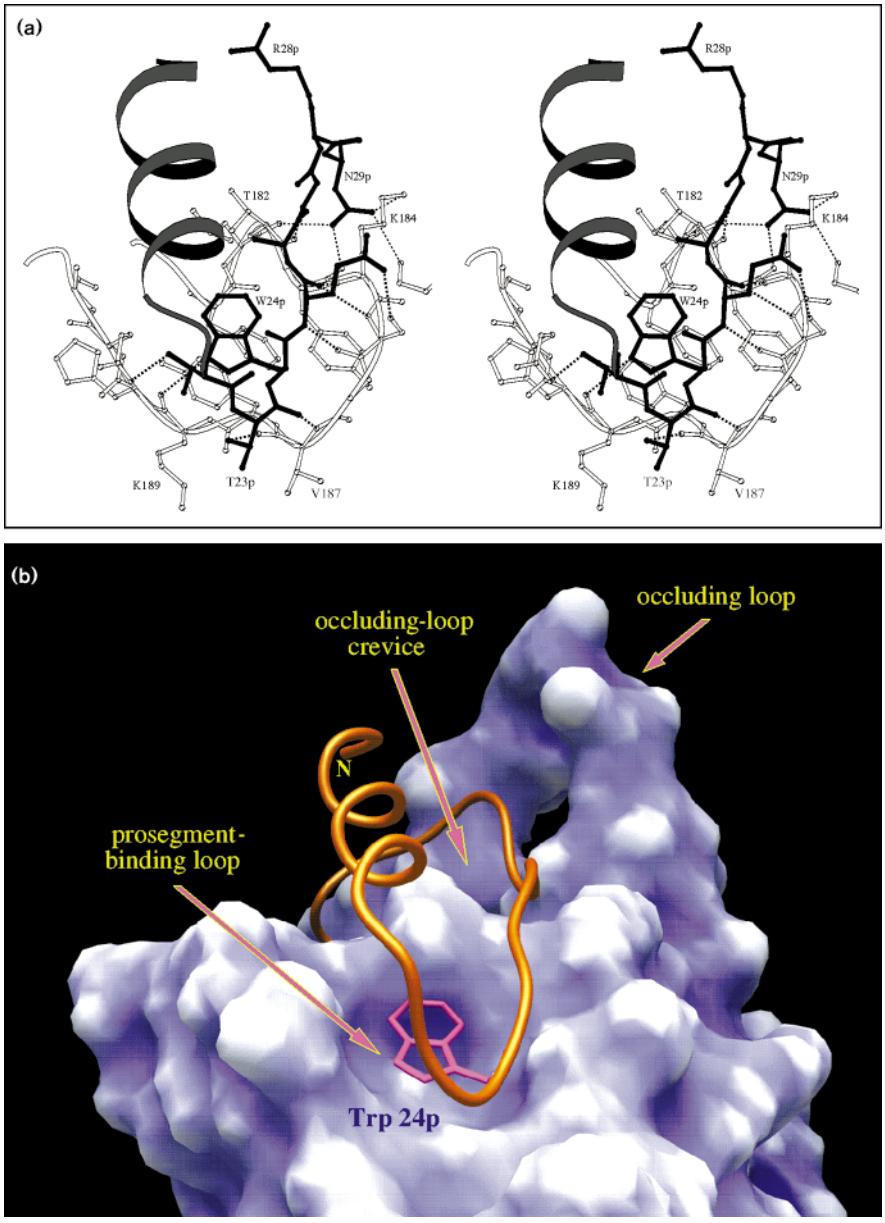


All-atom representation of the prosegment. α helices and β sheets are marked. **(a)** Stereoview of a ball-and-stick representation of the prosegment. **(b)** The N-terminal hairpin: side chains forming the hydrophobic core are drawn in dark lines. (The figure was prepared using MOLSCRIPT [47].)

prosegment interacts with PBL via the hairpin element, directing the β1p strand towards one side of this loop and the α1p helix towards the other. This orientation positions the side chain of Trp24p directly in the center of the depression, in such a way that it forms an additional member of the above mentioned aromatic cluster (Fig. 4). This residue alone provides ~100 Å^2 of contact area. In addition, the Nε1 of this tryptophan forms a hydrogen bond to the carbonyl of Lys189. The β1p strand forms a short, distorted β sheet with an extended Thr182–Val187 stretch with five hydrogen bonds between the main chain atoms (Fig. 4a). Additionally, the side chain of Thr23p is also hydrogen bonded to the carbonyl of Val187. The association between the hairpin and catB shields many hydrophobic residues from the solvent. Apart from Trp24p, these include Met13p (45 Å^2 of buried surface), Ile17p (39 Å^2), Thr23p (65 Å^2, hydroxyl group pointing out) and Ala26p (16 Å^2). The total surface area of the 12p–28p fragment buried by association with catB is 451 Å^2. The corresponding buried area of catB is 446 Å^2. Propeptides truncated at the N terminus provided valuable insight into the contribution of the N-terminal residues to the inhibitory properties of the propeptide. It was shown that a removal of up to twenty residues has relatively little effect on the K_i value (T Fox, E de Miguel, SV Nymark, JSM and ACS, unpublished data; ACS, unpublished data). Similarly, it was shown that Trp24p plays a pivotal role in binding of the hairpin. Its replacement by alanine increased K_i by more than two orders of magnitude (ACS, unpublished data).

A comparison with the structure of rat catB [26] shows that the removal of the proregion during the activation step has a

Figure 4



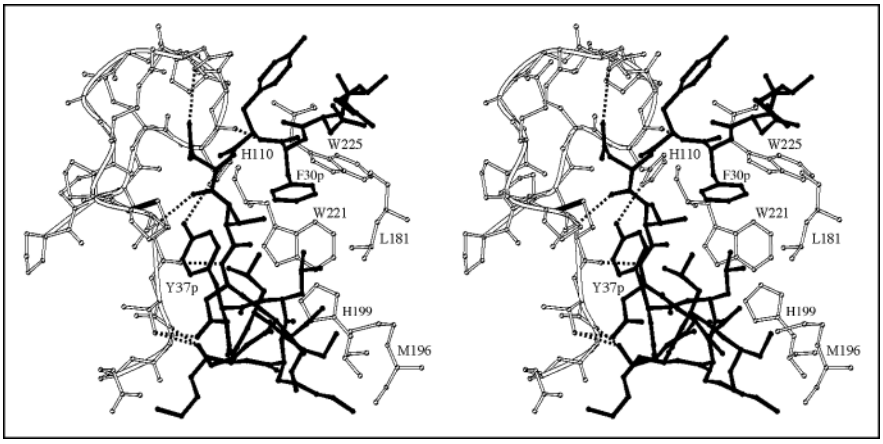
Interactions of the N-terminal hairpin with the prosegment-binding loop (PBL, residues 176–194). Hydrogen bonds are shown as dashed lines. **(a)** A detailed view, prepared using MOLSCRIPT [47]. **(b)** GRASP created surface of the PBL with the prosegment shown as a backbone worm. The side chain of Trp24p, centrally located on the PBL surface, is shown in full.

minimal effect on the conformation of the PBL. The differences are only small shifts (less than 1 Å) of the main chain for residues 189–194, a rotation of Glu191 side chain and ~30° rotation of the aromatic ring of Tyr188 to facilitate interaction with Trp24p. In the crystals of mature catB some ordered solvent molecules have been identified near the surface that in procathepsin B forms an interface with the proregion hairpin. Very few such molecules are near the center of PBL, reflecting a rather hydrophobic character of this region. Interestingly, solvent molecules in the catB structure were located close to the positions taken by the carbonyl oxygen of Ala25p, the hydroxyl of Tyr16p and the tip of the Asn29p side chain in the structure of procathepsin B.

Occluding-loop crevice

The second important area of interaction is the crevice formed by the occluding loop (108–122) on one side, part of PBL on the other side, and the 221–226 loop at the base (Fig. 5). The tip of the occluding loop is more mobile than the rest of the loop and the side chain of Arg116 is disordered. At Tyr31p–Asn32p, the prosegment changes direction by almost 90° and enters the interface between the two domains of catB. Residues 29p–40p, which include most of the α2p (35p–41p) helix, are wedged between the occluding loop and residues 176–184 from the PBL loop. In addition, the β strand 196–199 also participates in these contacts. There are eight hydrogen bonds between the

Figure 5



Contacts between the prosegment and catB within the occluding-loop crevice. Hydrogen bonds are shown as dashed lines. The prosegment is drawn in dark lines. (Figure was created using MOLSCRIPT [47].)

prosegment and the occluding loop. Three of them are between main chain atoms and occur at opposite ends of the occluding loop: $N^{Tyr31p}...O^{His110}$, $O^{Asn32p}...N^{Cys119}$ and $N^{Asp34p}...O^{Cys119}$. The remaining are between the side chains of Tyr37p and His110, between Asn32p and Asn13, and between Thr120 and two other groups: Asp34p and Ser36p. His111, which appears to play an essential role in the exopeptidase activity of catB [25] (JSM, unpublished data), is located on the solvent exposed side of the loop and points away from the active site. Two aromatic side chains of the prosegment, Phe30p (61 Å^2) and Tyr37p (87 Å^2), are buried in a hydrophobic environment near the bottom of the crevice and form an aromatic cluster together with His110, Trp221 and Trp225 of catB. The rings of His110 and Phe30p are nearly coplanar, with some degree of stacking between them. Val33p (22 Å^2) and Leu38p (45 Å^2), as well as Leu41p (77 Å^2) from the last turn of the α2p helix, contribute to this hydrophobic environment. The total buried surface is ~610 Å^2 for this part of the prosegment and the corresponding surface on catB is ~710 Å^2. The two basic residues, Lys39p and Lys40p, are on the surface of the protein and are exposed to the solvent. They make weak hydrogen bonds to the carbonyl groups of nearby residues of catB.

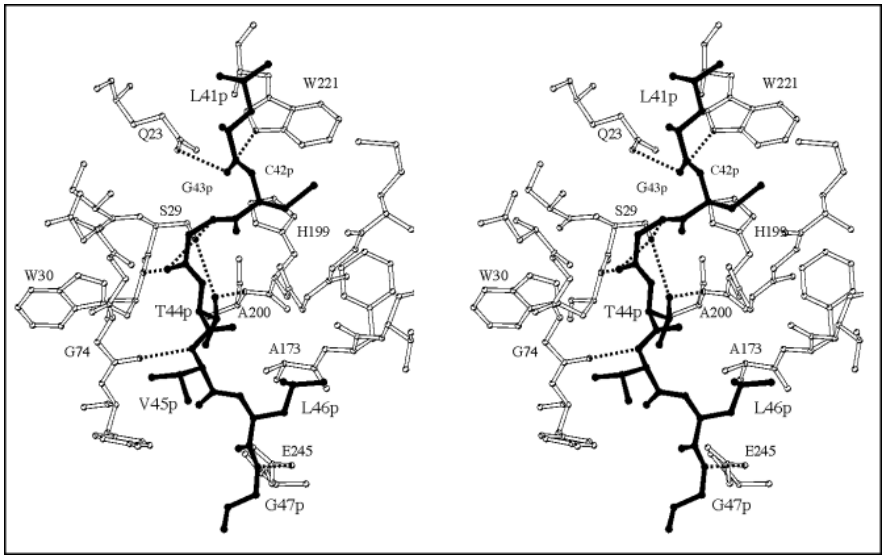
Substrate-binding cleft

The third major contact region is within the substrate-binding cleft and involves residues Leu41p to Gly47p of the proregion. They adopt an extended conformation and contact residues from the bottom and sides of the cleft (Fig. 6). This part of the proregion is largely shielded from the solvent; only 25% of its surface is solvent accessible. The direction of the polypeptide chain is opposite to that expected for natural substrates, as extrapolated from the structures of various covalent peptidyl inhibitors bound to catB [25,26] and papain [27] and implied from observed specificity changes caused by mutations of the S2-subsite residues [29,30]. This 'reverse' binding mode has only been observed previously in the metalloprotease prostromelysin [23], whereas for all other proteases whose proform structures are known ([20–22], for example) the prosegments occupy the substrate-binding sites in the same orientation as natural substrates. This ability to bind polypeptides to the same binding site in two opposite orientations is not unique to the active sites of proteases but has also been reported for SH3 domains [31].

Although the prosegment chain follows the substrate-binding cleft backwards, the S1, S2 and S1′, S2′ subsites, identified previously for catB [24–26], are nevertheless occupied by the prosegment's side chains in a manner that corresponds to the assigned subsite specificity. Leu41p (77 Å^2 buried area) extends into the S2′ subsite and contacts residues Trp221 and His110 from catB. Its carbonyl oxygen is hydrogen bonded to the oxyanion-hole-forming residue, Gln23, and to Nε1 of Trp221. Subsite S1′ is occupied by the side chain of Cys42p (60 Å^2) which interacts with many hydrophobic residues: Val176, Leu181, Met196, Gly197, Gly198 and Trp221. The carbonyl group of this cysteine is the closest to the oxyanion hole but the oxygen points away from it (Fig. 6). This side chain has a significant contribution to the binding of a propeptide. Its replacement by alanine or serine results in two orders of magnitude increase in K_i (ACS, unpublished data). The residue nearest to Ser29 (active site) is Gly43p. This residue penetrates deeply into the binding pocket, forms a hydrogen bond to the main chain NH of Trp30 and is also in close contact with Gly27. As we used a Cys29→Ser mutant, we observe additional weak hydrogen bonds from $O\gamma^{Ser29}$ to O^{Gly43p} and N^{Gly43p}. Replacing Gly43p with any other residue requires changes in the position of the prosegment in this region because of the close contacts of the Cα atoms of Gly43p and Gly27. Thr44p occupies the S2 subsite known to be the major specificity determinant for the proteases of the papain family. This residue is in the vicinity of Ala200, the main chain of His199 and Cβ of Trp30 and its Oγ forms

Figure 6

Interactions between the prosegment and catB within the substrate-binding cleft. Dashed lines indicate hydrogen bonds. The prosegment is shown in dark lines.



hydrogen bonds with $O\gamma^{Ser29}$ and N^{Ala200}. Thr44p fills only part of the available space, consistent with the enzyme's preference for a substrate with either a phenylalanine or an arginine in this position. Val45p points toward the solvent and fits into the crevice between the Tyr75 side chain and Gly73–Gly74 and its NH group forms a hydrogen bond with the carbonyl of Gly74. The next residue, Leu46p, occupies a shallow pocket delimited by Gly198, the main chain of Phe174 and Thr175 and the aliphatic part of the Glu245 side chain. Finally, Gly47p emerges on the surface of catB and forms a hydrogen bond with the side chain of Glu245. The latter residue is known to form a salt bridge with those substrates that contain an arginine in the P2 subsite [26]. The total buried area of this segment is 400 Å^2.

The oxyanion hole, flanked ordinarily by the NH of Cys29 and Nϵ2 of Gln23, which both form hydrogen bonds to the oxyanion, is left empty in the proenzyme. The N^{Ser29} does not form any hydrogen bonds, but the $N\epsilon2^{Gln23}$ forms two hydrogen bonds, to the main chain carbonyls of Lys40p and Leu41p, at the edge of the oxyanion hole.

Although, as indicated above, the propeptide contacts several residues lining the active site, modification of many of these positions, specifically Gln23, Gly73, His110 and Glu245, did not significantly effect inhibition by the propeptide (T Fox, E de Miguel, SV Nymark, JSM and ACS, unpublished data).

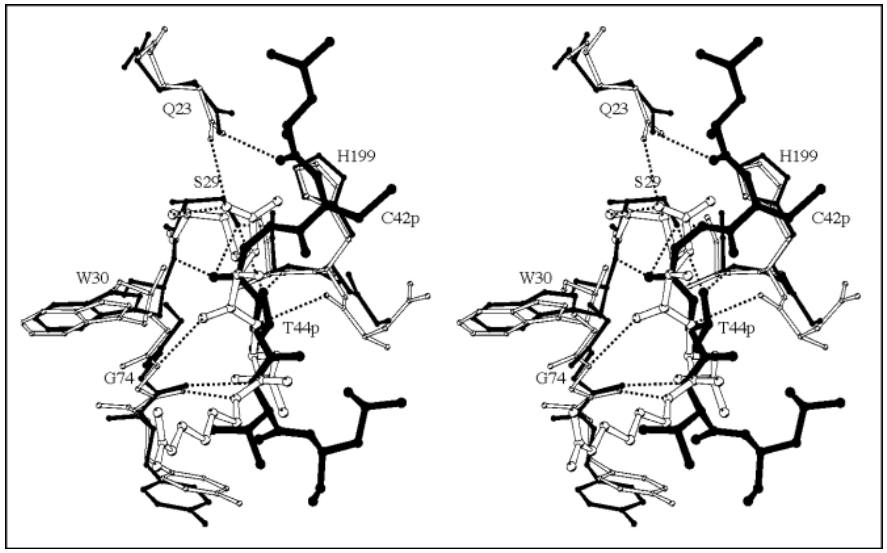
C terminus of the prosegment.

Beyond residue 48p the prosegment lies on a relatively flat surface of catB and makes fewer contacts with the enzyme. This part of the prosegment seems to act mostly as a physical linker between the tight-binding portion of the prosegment and catB. It is more mobile, particularly residues 55p–61p, than the central part of the prosegment and does not contribute significantly to the specificity of interactions. The hydrogen bonds with catB are formed mainly by the main chain atoms of residues 51p to 54p. Of the charged side chains, only Arg54p forms a weak salt bridge (to Glu163). The side chains of Glu59p and Asp60p face the solvent and are disordered. The relatively small contribution of the C-terminal portion to the inhibitory capabilities of a free propeptide was shown by the fact that the truncation from the C terminus to residue 46p had only a modest effect on the K_i (T Fox, E de Miguel, SV Nymark, JSM and ACS, unpublished data).

Comparison with the E-64 complexes

Of the various peptidomimetic inhibitors of papain-like cysteine proteases, a very potent but relatively non-selective inhibitor, E-64 (*L*-3-carboxy-2,3-*trans*-epoxypropionyl-leucyl-amino(4-guanidino)butane), is known to bind in the active site in a reverse orientation to that of natural substrates. As shown by the structure of procathepsin B, E-64 actually mimics the mode of binding of the prosegment. The structures of E-64 complexes with actinidin [32], papain [33] and a complex formed between an E-64 derivative, E-64c, and papain [34,35] suggest that these inhibitors bind to all cysteine proteases from the papain superfamily in a similar fashion. In Figure 7 we have superimposed actinidin–E-64 and procathepsin B to display relative positions in the substrate-binding site of the E-64 and a corresponding Cys42p–Val45p fragment of the prosegment. It is apparent that, although the side chains protrude into the same pockets, their interactions with the enzyme are not exactly the same. The two molecules are shifted within the binding site by approximately one bond length (~1.5 Å).

Figure 7



Superposition of the actinidin–E-64 complex and procathepsin B. This close-up of the substrate-binding cleft shows similarities in the binding mode between the E-64 inhibitor and the prosegment. Actinidin is shown in thin open lines, E-64 in thick open lines, catB in thin full lines and the prosegment in thick full lines. Hydrogen bonds are shown as dashed lines.

This is emphasized by the hydrogen bond pattern to an invariant residue in papain-like cysteine proteases, Gly74: E-64 forms two hydrogen bonds to Gly74 using the NH and C=O groups of the main-chain atoms flanking the leucine occupying the S2 site. These two hydrogen bonds to Gly74 also characterise substrate-like inhibitors despite a reversal of direction of the polypeptide chain [33]. In the proenzyme the S2 site is occupied by Thr44p. The equivalent hydrogen bond to the C=O group of Gly74 is formed by Val45p, whereas the carbonyl of Gly43p is shifted too far from the NH of Gly74 to form a hydrogen bond. Instead, it makes hydrogen bonds to the NH of Trp30 and to the Oγ of Ser29. The second significant difference between E-64 and the prosegment is that in the E-64 complex the oxyanion hole is occupied by a carbonyl oxygen but it is empty in procathepsin B. These differences may be related to a change of S2 subsite occupant: leucine in E-64 to threonine in the prosegment. The latter, Thr44p, which is involved in two hydrogen bonds, one of them to the Ser29 side chain, is shifted towards the edge of the S2 subsite, in the direction of the active site. Ser29 forms two additional hydrogen bonds to Gly43p. Whether the observed positioning of the prosegment in the binding-site cleft is affected by the Cys29→Ser mutation or reflects the known fact that for the binding in the reverse mode the type of side chain occupying the S2 subsite is less crucial [36], remains unclear. It appears, however, that binding in the reverse mode allows more flexibility than binding in the substrate-binding mode.

Conformational changes during proenzyme activation

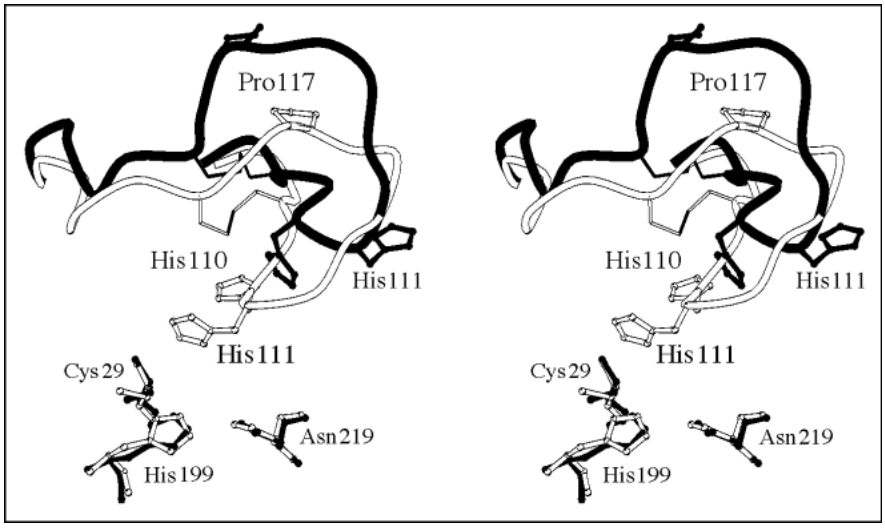
Occluding loop

Comparison of the structure of mature rat catB [26] with the equivalent portion of the procathepsin B provides information about the nature of conformational changes that occur upon activation. The examination of distances between the equivalent atoms after superposition of the two molecules shows that significant differences only occur within the 105–125 region. This is the occluding loop, characteristic of catB and a small number of proteins from the same subfamily but not present in the majority of cysteine proteases [17]. When the prosegment is removed in the course of the activation process approximately 400 Å^2 of the occluding loop surface becomes exposed and the structural support for this loop largely disappears. The loop refolds to a conformation observed in the mature catB (Fig. 8). The middle of the loop is fastened by a disulfide bridge, Cys108–Cys119. The cysteines act as main pivot points for the motion of the loop, which can be best described as twisting the top portion of the loop around an axis parallel to the strands of the bottom part of the loop. This motion shifts the main chain of the central part of the loop by as much as 10 Å. The largest displacement is that of the side chain of His111, the tip of which moves over 14 Å, with a simultaneous large movement of His110.

The occluding loop is crucial for the exopeptidase activity of catB. A mutant with this loop deleted shows nearly wild-type level of endopeptidase activity but greatly diminished exopeptidase activity (JSM, unpublished data). Within this loop the two histidines, His110 and His111, were predicted to be involved in binding of the C-terminal carboxylic group of the substrate [25]. This was based on the crystal structure of the complex of human cathepsin B and the CA030 inhibitor, which was proposed to simulate the binding of an exopeptidyl substrate. The critical role of the histidines for the exopeptidase activity has been confirmed by site-directed mutagenesis (JSM, unpublished data).

Figure 8

Comparison of the conformation of the occluding loop in procathepsin B (filled lines) and in the mature enzyme (open lines). Catalytic triad residues are shown. His110, His111 and Pro117 side chains are shown in full in medium lines and the Cys108–Cys119 disulfide bridge is shown in thin lines.



The hydrolysis of a small peptidyl substrate may proceed without reorganization of the occluding loop in the mature enzyme; however, the structure of procathepsin B suggests that for a protein substrate to be able to approach the active site the occluding loop has to move away. A similar hypothesis was postulated previously with regard to binding of cystatins to catB, namely that it has to be associated with a conformational change of the occluding loop [24]. If this is the case, does the occluding loop in the cystatin–catB complex assume the same conformation as that observed in the proenzyme structure? To answer this question one has to analyze the interactions that stabilize the observed conformations. The 'closed' conformation observed in catB forms many such contacts. His110 stacks against Trp221 and makes a hydrogen bond with Asp22; His111 is in close contact with Leu181, and Val112 with the backbone of Asp224 and Trp225. Additionally, Arg116 forms a salt bridge with Asp224. There are also many contacts between residues within the loop itself, creating a compact structure. In contrast, in the 'open' conformation observed in procathepsin B the structure of the top portion of the loop is much less compact and there are only a few contacts to catB, made by His110 and His111. The interactions that stabilize the observed open conformation are those with the prosegment; the shape of the occluding loop reflects the shape of the prosegment rather than an internally stable structure of the loop itself. While the conformation of the occluding loop in catB represents a conformation stabilized by interactions with the rest of the protein and relevant for its exopeptidase activity, that in the proenzyme is likely to represent one of many possible conformations that can be adopted by this loop with protein substrates. Thus, there seems to be no support for the hypothesis that there is only one open conformation of the occluding loop; rather, binding of the substrate causes switching from a single closed conformation to an open state in which the exact conformation depends on the substrate itself. Reorganization of this loop during substrate binding is likely to consume some of the binding energy and may account for the low endopeptidase activity of catB. This is supported by the findings that the inhibition of catB by peptides corresponding to parts of the prosegment is adversely affected by the presence of the occluding loop. When this loop is removed the K_i decreases by one order of magnitude (ACS, unpublished data) which indicates that forcing the 'opening' of the occluding loop is costly in energetic terms and that this is not offset by the interactions with the propeptide.

N terminus

Crystallographic studies of the proforms of various proteases showed that in the maturation process, during which the prosegment is removed, the N terminus of the mature enzyme very often reorganizes and moves from a few to over 40 Å [11,21,23]. Papain-like cysteine proteases from the cathepsin B subfamily display a different behaviour. Comparison of the structures of procathepsin B and catB clearly shows that the catB N terminus is in the same position as observed in the proenzyme. Thus, activation and cleavage at the amino side of Leu1 does not lead to a rearrangement near the cleavage site and does not affect the conformation of the polypeptide chain. In this respect, procathepsin B is similar to zinc procarboxypeptidases A and B: activation also does not disturb the conformation of the newly created N terminus in these enzymes [37]. As the superposition of 3D structures of other cysteine proteases from the papain superfamily shows that positions and conformations of their N termini are well conserved, this lack of rearrangement may be characteristic not only for the cathepsin B subfamily but also for other papain-like proteins.

Mechanism of activation

The structure of the procathepsin B provides a basis for speculation on the activation mechanism. The C-terminal portion of the proregion, residues p54–p62, is well exposed to the solvent and is significantly more mobile (higher B factors) than the other parts of the proregion (with the exception of the N terminus). Some of the side chains in this C-terminal region are also poorly defined. Such an organization provides a processing protease with easy access to this region. Indeed, *in vitro* processing of the Cys29→Ser inactive procathepsin B mutant by various proteases showed that the cleavage occurs in this region, the precise site reflecting the specificity of the particular protease [15]. Although the cleavage can occur at neutral pH, the propeptide does not dissociate until the pH is lowered.

The activation, both *in vivo* and *in vitro*, is triggered by acidic pH. A simplistic explanation — lowering the pH causes protonation of some acidic groups with a concomitant weakening of salt bridges and/or hydrogen bonds between the prosegment and catB — is not supported by the structure of procathepsin B. Out of 35 hydrogen bonds between the prosegment and catB, 13 are of C=O...HN type and only five involve acidic groups, some of which could still be maintained in a protonated state. The prosegment contains seven acidic residues: Asp2p, Asp11p, Asp12p, Asp34p, Glu53p, Glu59p and Asp60p. Aspartate side chains are on the surface of the protein and make no hydrogen bonds. As the 6p–51p propeptide shows similar pH dependence of inhibition (T Fox, E de Miguel, SV Nymark, JSM and ACS, unpublished data) to the whole of the prosegment, three of these acidic groups cannot contribute significantly to the observed behavior. Furthermore, the replacement of remaining aspartates by asparagines, in a synthetic propeptide, did not modify the pH dependence of its binding constant relative to the wild-type sequence (T Fox, E de Miguel, SV Nymark, JSM and ACS, unpublished data). Glu53p does not form a salt bridge and it seems unlikely that this side chain is responsible for the pH sensitivity. As no specific interactions can be pinpointed it seems likely that the observed effect results from many small contributions. We would like to postulate the following picture of the activation. The role of the acidic pH is to loosen up the structure of the protein and increase the mobility of the prosegment. The first cleavage occurs in the 55p–62p region, and the loosely dangling ends are clipped away in subsequent steps. The now C-terminal part of the prosegment blocking the active-site liberates itself and is removed by cleavage. The binding of the remaining peptide becomes progressively less tight, eventually leading to the peptide falling off the enzyme. This hypothesis can not be validated solely from crystallographic data. NMR spectroscopy provides ideal tools for this purpose and will be employed to study further the pH-triggered activation process.

Selectivity of the propeptide

Biochemical data on the specificity of the catB propeptide is rather limited. It has been shown that a synthetic propeptide, corresponding to the 1p–56p fragment of procathepsin B, inhibits catB with K_i=0.4 nM but is ineffective against cathepsin L, S and papain (K_i in μM range) ([7]; T Fox, E de Miguel, SV Nymark, JSM and ACS, unpublished data). Similarly, although the propeptide corresponding to the prosegment of procathepsin L is a nanomolar inhibitor of cathepsin L, it inhibits catB at least 10^4 times less efficiently; (E Carmona *et al.*, & R Ménard, unpublished data). However, it remains to be determined if the proregions show specific recognition within the family of closely related catB-like proteases. Without a detailed knowledge of the structure of the proregion from members of the papain superfamily with longer prosegments (e.g. propapain, procathepsin L), it is rather speculative to try to identify the key factors that determine the specificity of the prosegment. However, if there is any similarity in the binding mode of the papain prosegment (relative to that of catB) in the active-site cleft, the prosegment will interact with the area corresponding to the occluding-loop crevice of catB. A comparison of these two proteases shows two significant structural differences in that region. The first is within the occluding loop, which has a long insertion in catB-like proteases; the second is in a very different path of the residues, 190–198 (catB), the rim of the PBL loop. Although the occluding loop is not essential for binding, as was shown by the occluding-loop-free mutant (JSM, unpublished data), the proregion of procathepsin B interacts with all of the 190–198 residues. Importantly, as they form one side of the occluding-loop crevice, they also determine the surface shape to which the prosegment has to be complementary. It is then likely that the 190–198 region is an important element that differentiates between these two subfamilies.

Biological implications

Cysteine proteases belonging to the papain superfamily are important components of the lysosome and are involved in intracellular protein turnover. They have also been implicated in extracellular matrix protein degradation associated with many pathological conditions, which makes them important targets for inhibitor development. Like most proteases, they are synthesized as inactive precursors with N-terminal extensions that are potent and highly specific inhibitors of their enzymatic activity. Although inhibitors with sufficient potency are known, they lack the required specificity for therapeutic applications. The structure of procathepsin B provides a detailed picture of the prosegment–enzyme interactions and the mechanism of inhibition and should stimulate new ideas for designing inhibitors with enhanced specificity.

As is the case for many other proproteases, the prosegment disables the enzymatic activity by extending along the substrate-binding site and blocking access to the active site. The distinguishing feature of procathepsin B when compared with other proteases, with the exception of some metalloproteases, is a reverse direction of the prosegment in the substrate-binding site relative to that taken by the substrate. This is similar to the mode of binding of E-64, a potent, naturally occurring small inhibitor of cysteine proteases, which rather than following the substrate-binding mode mimics the binding mode of the prosegment. The prosegment has no globular structure on its own and folds around catB, interacting with three main areas: the substrate-binding cleft, the occluding-loop crevice and the prosegment-binding loop (PBL). The PBL provides a large, exposed, partly hydrophobic surface that is enlisted by the prosegment and used as an auxiliary binding site. Such an extended area of interaction augments the specificity of the prosegment for its parent protease.

The utilization of a secondary site, remote from the substrate-binding site, to significantly improve binding and/or to augment the specificity of recognition has been observed in other situations e.g. in cases of small protein inhibitors binding to their targets. The well studied example is that of hirudin binding to thrombin [38] in which hirudin utilizes a secondary fibrinogen binding exosite in addition to the substrate-binding site to achieve extremely tight binding. This bi-functional binding mechanism has been successfully exploited in the design of potent inhibitors of thrombin [39]. Similar approaches should lead to development of potent and specific inhibitors of cysteine proteases.

Materials and methods

Rat procathepsin B was expressed in the methylotrophic yeast *Pichia pastoris* as an α-factor fusion construct in vector pPIC9 (Invitrogen Inc., San Diego, California) using the cDNA insert previously expressed in *Saccharomyces cerevisiae* [15]. As the wild-type enzyme easily undergoes self-processing, for crystallographic studies a mutation of the active site Cys29 to serine was constructed. A second mutation, Ser115→Ala, modified the N-glycosylation site NGS^{115} within the occluding loop (108–122) to NGA^{115}. That left only one N-linked glycosylation site, on Asn21p of the prosegment. Sequencing the cDNA showed that an additional difference from the published sequence, Ala223 in place of Val, was present. This is likely to be a naturally occurring polymorphism as alanine was also observed in the original clone that was used in this work [40]. This replacement has no effect on the activity of a wild-type catB, as was confirmed by reversing this mutation (JSM, unpublished data).

Recombinant procathepsin B has been crystallized [41] in the trigonal space group $P3_121$ with cell dimensions a=99.6, c=141.4 Å, γ=120° and contains two molecules in the asymmetric unit. The native data set was collected at room temperature to 2.8 Å resolution on an R- Axis IIC area detector placed on an RU-300 rotating anode generator. The completeness of the data was 91%. The last data shell, 2.9–2.8 Å, contained rather weak reflections, $\langle I/\sigma(I)\rangle$=1.7. The coordinates of mature rat catB [26] (PDB code 1cte) were taken as a starting model. The rotation function calculated in the resolution range 8–4 Å and with the radius of 25 Å using the program AMoRe [42] had two peaks significantly above background of nearly equal heights. The translation function confirmed that each of these peaks corresponded to an orientation of one of the two molecules in the asymmetric unit; in each case a clear, unique solution was found with a correlation coefficient of 0.30 and an R factor of 0.46 for 8–3.5 Å data. Combining the two solutions resulted in a correlation coefficient of 0.56 and an R factor of 0.374. Rigid-body refinement performed with AMoRe increased the correlation coefficient to 0.60 and reduced the R factor to 0.358. Further refinement was conducted with the program X-PLOR including data with $I>\sigma(I)$ in the 8–2.8 Å resolution range. The test set for calculations of R free [43] consisted of 10% of randomly selected reflections.

The electron-density map calculated at that stage clearly showed part of the prosegment near the active site, but the density for the remaining part was unclear. This map indicated, however, a general location of the proregion and allowed the construction of a mask for the whole molecule. Several cycles of solvent flattening coupled with non-crystallographic symmetry averaging (RAVE [44]) improved significantly the electron-density map and most of the proregion could be traced. The model rebuilding was done with program O [45]. The occluding loop, which was removed from the original catB model, was built in a new conformation into well-defined density, with the exception of the tip of the loop where the side chain of Arg116 could not be located. During all stages of refinement the non-crystallographic symmetry restraints were applied. After one round of minimization the remainder of the proregion, with the exception of the N terminus, was built. This model was refined with all data (2.8 Å). Although the electron density at the end of the carbohydrate-bearing Asn21p side chains is ambiguous, there is a sufficient space around them in both molecules to accommodate the carbohydrate. The residues included in the model are 10p to 62p and 1 to 260. The current R factor for all data is 0.22 and R free is 0.27 (for data with $I>2(I)$ the R-factors are 0.19 and 0.24, respectively). The rms deviations for bond lengths and angles are 0.009 Å and 1.44°, respectively. The model has been additionally checked with the program PROCHECK [46] which showed that only one residue, Asn 222, is in a disallowed region of the Ramachandran plot. This residue in the structure of catB [26], determined at 2.2 Å resolution, has similar ϕ,ψ angles.

The coordinates of rat procathepsin B have been deposited with the Brookhaven Databank, accession number 1min.

Acknowledgements

We gratefully acknowledge the expert technical assistance of Patrizia Mason and Marie-Claude Magny in protein preparation and guidance of Dr Svetlana Borisova and Yunge Li during crystallization experiments. NRCC publication No. 39913.

References

1. Knop, M., Schiffer, H.H., Rupp, S. & Wolf, D.H. (1993). Vacuolar/lysosomal proteolysis: proteases, substrates, mechanisms. *Curr. Opin. Cell Biol.* **5**, 990–996.
2. Takahashi, H., Cease, K.B. & Berzofsky, J.A. (1989). Identification of proteases that process distinct epitopes on the same protein. *J. Immunol.* **142**, 2221–2229.
3. Vaes, G. (1988). Cellular biology and biochemical mechanisms of bone resorption. A review of recent developments on the formation, activation and mode of action of osteoclasts. *Clin. Orthop.* **231**, 239–271.
4. Sloane, B.F. & Honn, K.V. (1984). Cysteine proteinases and metastasis. *Cancer Metastasis Rev.* **3**, 249–263.
5. Esser, R.E., *et al.*, & Smith, R.E. (1994). Cysteine proteinase inhibitors decrease articular cartilage and bone destruction in chronic inflammatory arthritis. *Arthritis Rheum.* **37**, 236–247.
6. Neurath, H. (1989). Proteolytic processing and physiological regulation. *Trends Biochem. Sci.* **14**, 268–271.
7. Fox, T., de Miguel, E., Mort, J.S. & Storer, A.C. (1992). Potent slow-binding inhibition of cathepsin B by its propeptide. *Biochemistry* **31**, 12571–12576.

8. Taylor, M.A.J., *et al*., & Goodenough, P.W. (1995). Recombinant pro-regions from papain and papaya proteinase IV are selective high affinity inhibitors of the mature papaya enzymes. *Protein Eng.* **8**, 59–62.
9. Tao, K., Stearns, N.A., Dong, J., Wu, Q. & Sahagian, G.G. (1994). The proregion of cathepsin L is required for proper folding, stability, and ER exit. *Arch. Biochem. Biophys.* **311**, 19–27.
10. Mach, L., Mort, J.S. & Glössl, J. (1994). Non-covalent complexes between the lysosomal proteinase cathepsin B and its propeptide account for stable, extracellular, high molecular mass forms of the enzyme. *J. Biol. Chem.* **269**, 13036–13040.
11. Gallagher, T., Gilliland, G., Wang, L. & Bryan, P. (1995). The prosegment-subtilisin BNP' complex: crystal structure of a specific 'foldase'. *Structure* **3**, 907–914.
12. Smith, S.M. & Gottesman, M. (1989). Activity and deletion analysis of recombinant human cathepsin L expressed in *Escherichia coli*. *J. Biol. Chem.* **264**, 20487–20495.
13. Vernet, T., *et al*., & Thomas, D.Y. (1991). Processing of papain precursor. *J. Biol. Chem.* **266**, 21451–21457.
14. Mach, L., Mort, J.S. & Glössl, J. (1994). Maturation of human cathepsin B. Proenzyme activation and proteolytic processing of the precursor to the mature proteinase, *in vitro*, are primarily unimolecular processes. *J. Biol. Chem.* **269**, 13030–13035.
15. Rowan, A.D., Mason, P., Mach, L. & Mort, J.S. (1992). Rat procathepsin B. Proteolytic processing to the mature form *in vitro*. *J. Biol. Chem.* **267**, 15993–15999.
16. Kirschke, H. & Barrett, A.J. (1987). Chemistry of lysosomal proteases. In *Lysosomes: Their role in protein breakdown*. (Glaumann, H. & Ballard, F.J., eds), pp. 193–238, Academic Press, London.
17. Berti, P. & Storer, A.C. (1995). Alignment/phylogeny of the papain superfamily of cysteine proteases. *J. Mol. Biol.* **246**, 273–283.
18. Karrer, K.M., Peiffer, S.L. & DiTomas, M.E. (1993). Two distinct gene subfamilies within the family of cysteine protease genes. *Proc. Natl. Acad. Sci. USA* **90**, 3063–3067.
19. Huber, R. & Bode, W. (1978). Structural basis of the activation and action of trypsin. *Accounts Chem. Res.* **11**, 114–122.
20. Bryan, P., *et al*.,. & Gallagher, T. (1995). Catalysis of a protein folding reaction: Mechanistic implications of a 2.0 Å structure of the subtilisin–prodomain complex. *Biochemistry* **34**, 10310–10318.
21. James, M.N.G. & Sielecki, A.R. (1986). Molecular structure of an aspartic proteinase zymogen, porcine pepsinogen, at 1.8 Å resolution. *Nature* **319**, 33–38.
22. Coll, M., Guasch, A., Avilés, F.X. & Huber, R. (1991). Three-dimensional structure of porcine procarboxypeptidase B: a structural basis of its activity. *EMBO J.* **10**, 1–9.
23. Becker, J.W., *et al*., & Springer, J.P. (1995). Stromelysin-1: Three-dimensional structure of the inhibited catalytic domain and of the C-truncated proenzyme. *Prot. Sci.* **4**, 1966–1976.
24. Musil, D., *et al*., and Bode, W. (1991). The refined 2.15 Å X-ray crystal structure of human liver cathepsin B: the structural basis for its specificity *EMBO J.* **10**, 2321–2330.
25. Turk, D., *et al*., & Turk, V. (1995). Crystal structure of cathepsin B inhibited with CA030 at 2.0 Å resolution: A basis for the design of specific epoxysuccinyl inhibitors. *Biochemistry* **34**, 4791–4797.
26. Jia, Z., *et al*., & Huber, C.P. (1995). Crystal structures of recombinant rat cathepsin B and a cathepsin B–inhibitor complex: implications for structure-based inhibitor design. *J. Biol. Chem.* **270**, 5527–5533.
27. Baker, E.N. & Drenth, J. (1987). The thiol proteases: Structure and mechanism. In *Biological Macromolecules and Assemblies*, Vol. **3**. (Jurnak, F.A. & McPherson, A. eds), pp. 313–368, John Wiley & Sons, New York.
28. Moore, S.A., Sielecki, A.R., Chernaia, M.M., Tarasova, N.I. & James, M.N.G. (1995). Crystal and molecular structures of human progastricsin at 1.62 Å resolution. *J. Mol. Biol.* **247**, 466–485.
29. Hasnain, S., Hirama, T., Huber, C.P., Mason, P. & Mort, J.S. (1993). Characterization of cathepsin B specificity by site-directed mutagenesis. Importance of Glu245 in the S2-P2 specificity for arginine and its role in transition state stabilization. *J. Biol. Chem.* **268**, 235–240.
30. Khouri, H.E., *et al*., & Storer, A.C. (1991). Engineering of papain: selective alteration of substrate specificity by site-directed mutagenesis. *Biochemistry* **30**, 8929–8936.
31. Feng, S.B., Chen, J.K., Yu, H.T., Simon, J.A. & Schreiber, S.L. (1994). Two binding orientations for peptides to the src SH3 domain – development of a general model for SH3–ligand interactions. *Science* **266**, 1241–1247.
32. Varughese, K.I., Su, Y., Cromwell, D., Hasnain, S. & Xuong, N.H. (1992). Crystal structure of an actinidin–E-64 complex. *Biochemistry* **31**, 5172–5176.
33. Varughese, K.I., *et al*., & Storer, A.C. (1989). Crystal structure of papain–E-64 complex. *Biochemistry* **28**, 1330–1332.
34. Yamamoto, D., *et al*., & Mizuno, H. (1991). Refined x-ray structure of papain–E-64-c complex at 2.1 Å resolution. *J. Biol. Chem.* **266**, 14771–14777.
35. Kim, M.J., *et al*., & Kitamura, K. (1992). Crystal structure of papain–E64-c complex. Binding diversity of E64-c to papain S2 and S3 subsites. *Biochem. J.* **287**, 797–803.
36. Gour-Salin, B.J., *et al*., & Storer, A.C. (1995). E-64 [trans-epoxysuccinyl-L-leucylamido-(4-guanidino)butane] analogues as inhibitors of cysteine proteinases: investigation of S2 subsite interactions. *Biochem. J.* **299**, 389–392.
37. Guasch, A., Coll, M., Aviles, F.X. & Huber, R. (1992). Three-dimensional structure of porcine pancreatic procarboxypeptidase A. A comparison of the A and B zymogens and their determinants for inhibition and activation. *J. Mol. Biol.* **224**, 141–157.
38. Rydel, T.J., *et al*., & Fenton, J.W. Jr. (1990). The structure of a complex of recombinant hirudin and human α-thrombin. *Science* **249**, 277–280.
39. DiMaio, J., *et al*., & Konishi, Y. (1990). Bifunctional thrombin inhibitors based on the sequence of hirudin. *J. Biol. Chem.* **265,** 21698–21703.
40. San Segundo, B., Chan, S.J. & Steiner, D.F. (1985). Identification of cDNA clones encoding a precursor of rat liver cathepsin B. *Proc. Nat. Acad. Sci. USA* **82,** 2320–2324.
41. Sivaraman, J., *et al*., & Cygler, M. (1996). Crystallization of rat procathepsin B. *Acta Cryst. D*, in press.
42. Navaza J. (1994). AMoRe: an automated package for molecular replacement. *Acta. Cryst.* **50**, 157–163.
43. Brünger, A.T. (1992). The free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **355**, 472–474.
44. Kleywegt, G.J. & Jones, T.A. (1994) Halloween... Masks and Bones. In *From first map to final model*. (Bayley. S., Hubbard R. & Waller D., eds), pp.59–66, SERC Daresbury Laboratory, Daresbury, England.
45. Jones, T.A., Zou, J.Y., Cowan, S.W. & Kjeldgaard, M. (1991). Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Cryst.* A**47**, 110–119.
46. Laskowski, R.A., MacArthur, M.W., Moss D.S. & Thornton J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* **26**, 283–291.
47. Kraulis, P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Cryst.* **24**, 946–950.
48. Nicholls, A., Sharp, K. & Honig, B. (1991). Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins* **11**, 281–296.